Meta's Approach to Disinformation and Misinformation: Contextualising Recent Developments

Dr Phoebe J Galbally
Lecturer, UWA Law School

On 7 January 2025, Meta's founder and CEO, Mark Zuckerberg, announced that Meta would cease third-party fact-checking of content on its platforms (including Facebook and Instagram), starting in the United States (US). In place of fact-checking, Zuckerberg announced that Meta would adopt a policy of flagging false speech using 'Community Notes', consistent with the approach of rival social media platform, X (formerly Twitter), owned by Elon Musk.[1] The main impetus for Meta's policy shift, which coincided with Donald Trump's second presidency, seems to be political. In this regard, Zuckerberg's announcement cited the 'recent elections' in the US as a 'cultural tipping point towards once again prioritising speech'.[2]

This 'cultural' shift may best be understood in the context of President Trump's social media communication being the subject of regulatory sanctions (including fact-checking) for spreading disinformation and misinformation. For example, by November 2020, X had attached warning labels to 38% of Trump's posts under its Civic Integrity Policy (addressing misleading electoral communication).[3] Further, and as a result of escalating violence that Meta deemed to be connected with his social media communication on Meta's platforms, Trump's accounts were suspended following the US Capitol Riots on January 6, 2021.[4]

Trump has since rallied against social media regulation, and on the first day of his second term as President, he signed an executive order 'ending federal censorship' of speech 'under the guise of combatting "misinformation" [and] "disinformation"'.[5] This executive order was issued even though the US has not enacted any Federal regulation of misinformation and disinformation. Moreover, the order simply restates the protection afforded by the First Amendment of the US Constitution, preventing federal officials from any conduct which 'would unconstitutionally abridge the free speech of any American citizen'.[6] Nevertheless, Trump's executive order is reflective of criticism that social media platforms have long faced, particularly from politically conservative quarters, that they are unduly censoring free speech.

In this article, I attempt to sort through the political rhetoric and explore Meta's policies on disinformation and misinformation. I begin with a brief overview of the approaches that may be taken when regulating social media misinformation and disinformation before examining Meta's approach to such regulation. This Article concludes by outlining Meta's policy changes and highlighting several key takeaways that may be drawn from these changes.

A Brief Overview of the Regulation of Social Media Disinformation and Misinformation

When it comes to regulating disinformation and misinformation, social media companies, such as Meta, largely follow a self-regulatory model in jurisdictions such as Australia and the US. This regulatory approach may be contrasted with external regulation which exists in the European Union (EU)[7] and the United Kingdom (UK).[8] External regulation may adopt a variety of methods for ensuring platform compliance with regulatory standards, ranging from voluntary commitments (such as those contained in the EU's Strengthened Code

of Practice on Disinformation of 2022) to mandatory legal obligations backed by sanctions (usually fines). While there are multifaceted reasons for the differences in the regulatory approaches states may take to address social media disinformation and misinformation, relevant considerations may be broadly grouped into two categories: political and legal.

Politically, as has been seen in Australia,[9] state regulation of misinformation and disinformation tends to be challenging to enact. In this regard, state regulation of disinformation and misinformation is met with common criticisms, including the difficulty in distinguishing truth from falsity in communication, the undesirability of the state (or a corporate entity) acting as an 'arbiter of truth', and legitimate concerns regarding the risk of chilling freedom of expression. In this context, policymakers must clearly articulate the harms sought to be addressed through the regulation of social media disinformation and misinformation. In the absence of a clear understanding of such harms, and consequently, the necessity for state action, it may be difficult to make a successful political argument for regulation.

Legally speaking, constitutional and human rights frameworks protecting the right to freedom of expression may also make it challenging to enact state regulation of social media disinformation and misinformation. For example, under the First Amendment to the US Constitution, speech may only be limited in a very narrow range of cases. With respect to disinformation, existing First Amendment jurisprudence protects intentionally false speech 'absent a particularised showing of social harm resulting from the speech'.[10] The First Amendment's highly speech-protective approach may be contrasted with the regulation of the right to freedom of expression in Europe. The European Convention on Human Rights, for example, is more permissive with respect to limitations on speech, such as speech amounting to Holocaust denial.[11]

An Overview of Meta's Self-Regulatory Approach: Content Moderation and the Oversight Board

Broadly speaking, there are two tiers to Meta's self-regulatory approach. The first tier of regulation is where the majority of decisions regarding content on Meta's platforms are made. This involves 'content moderation', which is where Meta determines, through human moderators and AI, what type of content and behaviour is permitted on its platforms, in accordance with its policies. This form of content moderation is the most common self-regulatory approach taken by social media platforms and is also employed by X.

A second tier of self-regulation occurs through Meta's 'Oversight Board'.[12] Meta's Oversight Board was announced in 2020 in response to increasing negative publicity regarding Meta's governance and decision-making accountability. The Oversight Board is a court-like body that presides over a very narrow selection of cases within its jurisdictional purview, and is designed to offer social media users the opportunity for independent review of content moderation decisions. However, given the sheer size of Meta platforms and the number of content-related decisions made by the company every day,[13] the Oversight Board suffers from a 'highly circumscribed' jurisdictional scope and its regulatory impact is therefore extremely limited.

Meta's Self-Regulation of Disinformation and Misinformation

Before examining Meta's self-regulatory approach to disinformation and misinformation, it is first necessary to define misinformation and disinformation. Definitions focus on intent in distinguishing disinformation from misinformation. Where a speaker has no intention to deceive, and the communication is disseminated accidentally or negligently, for example, it constitutes 'misinformation'. In contrast, 'disinformation' may be defined as the *intentional* dissemination of information which is 'untrue'.[14] This is reflected in the widely accepted 'European Union High Level Expert Group's' definition of 'disinformation' as 'false, inaccurate, or misleading information designed, presented or promoted to intentionally cause public harm or for profit'.[15]

Because a speaker *intends* to manipulate through speech, and as a consequence of the harms associated with 'disinformation', regulatory approaches addressing social media disinformation tend to be more comprehensive than those addressing misinformation. For example, in 2018, following investigations regarding foreign disinformation impacting the US 2016 Presidential election, Meta developed a Policy to address Coordinated Inauthentic Behavior (CIB). Meta's CIB Policy is directed at 'groups or Pages or people working to mislead others about who they are or what they are doing' to achieve a 'strategic' end.[16] Meta's CIB Policy therefore encompasses disinformation campaigns and provides for important transparency measures, including mandating monthly reporting as to the number of 'pages' and 'groups' Meta captures under the Policy, as well as any content that Meta removes in accordance with the Policy.[17] These reporting requirements have led to improvements in transparency and data on CIB on Meta's platforms, particularly via Meta's Adversarial Threat Report (issued quarterly).[18]

With respect to 'misinformation', Meta adopted three interrelated policy approaches: its former 'Third Party Fact-Checking Program', misinformation 'label's', and efforts to prevent 'misinformation' from 'going viral'.[19] These approaches worked to complement each other. Meta responded to independent fact-checker 'ratings' by limiting the ability for users to view content or by applying a label to such content. Meta's Third Party Fact-Checking Program therefore did not restrict or censor speech, but rather, moderated the visibility of such speech through the use of algorithms. In this regard, fact-checking policies are often seen as a preferable, speech-preserving regulatory response to certain forms of social media misinformation and disinformation. Independent fact-checking therefore features as an important initiative for platform Signatories to adopt as part of the EU's Strengthened Code of Practice on Disinformation of 2022.

Meta's former Policy, which relied on independent, third-party fact-checkers, also reflected industry best practice standards. Independent third-party fact-checkers were tasked with reviewing and rating the accuracy of content on Meta's platforms, by taking into account various forms of 'in-depth' data analysis, including analysis of photos and videos.[20] Meta's fact-checkers were certified through the International Fact Checking Network (IFCN), and adhered to the IFCN's code of principles (particularly rules regarding non-partisanship), as well as to 'commitments' contained in the EU's Strengthened Code of Practice on Disinformation of 2022.

The use of *independent* fact-checkers may be contrasted with fact-checking systems done 'in house' by a social media platform, relying on mainstream media reports to verify facts. For example, when Twitter (now X) issued its fact-check of a post by President Trump which stated that 'Mail in Ballots' used during the Covid 19 pandemic in the US 2020 Presidential election would result in electoral fraud and produce a 'Rigged Election', Twitter 'staff became targets for highly critical responses from President Trump and his supporters'.[21]

It has, therefore, been noted that the source of a 'fact-check' can be highly influential

in assessing its factual legitimacy, and the legitimacy users attribute to the fact-checking system itself.[22]

Despite the apparent virtues of Meta's Third Party Fact-Checking Program, in his January 2025 announcement, Zuckerberg described the Program as 'too politically biased' and as a 'tool to censor'.[23] However, Meta's lack of transparency and poor record of data disclosure makes objectively verifying such concerns unlikely.[24] Further, it is unclear how Meta's 'Community Notes' will improve on these issues. At this stage, Meta has simply stated that its 'Community Notes program' is 'a new way for the [Meta] community to decide when posts are potentially misleading and add more context'.[25]

As to how the 'program' will work, Meta has provided the following information. First, it will be 'Community Note contributors' (i.e. platform users) who will write and submit Community Notes to 'posts they think are potentially misleading, inaccurate or incomplete'.[26] Accordingly, and given that social media users are often influenced by political interests or 'ideological biases', Meta's 'Community Notes' program could also be impacted by 'political bias'.[27] Second, the system relies on user consensus in order to publish a Community Note. As Meta has explained, for a 'Community Note to be published … users who normally disagree based on how they've rated Notes in the past, will have to agree that a Note is helpful' for it to be published.[28] Third, Meta has stated that it intends 'to be transparent about how different viewpoints inform the Notes displayed', but is yet to provide any indication as to how such transparency will be achieved.[29] Finally, Meta has stated that it plans to 'phase in Community Notes in the United States first over the next couple of months, and will continue to improve it'.[30]

While much remains to be seen regarding Meta's 'Community Notes', we can draw several takeaways from Meta's policy shift. Through Trump's second presidency, and with Elon Musk holding a central role in the White House, issues concerning the right to freedom of expression and the regulation of social media communication are likely to continue to feature in the US political agenda. With platform governance in the spotlight, Zuckerberg has indicated his allegiance to Trump while announcing to Meta investors that 2025 'is going to be a big year for redefining our relationships with governments'.[31]

However, it is important to situate the policy changes made by Meta, as well as the political rhetoric surrounding social media regulation in the US, within the framework of the First Amendment of the US Constitution, which is exceptional in the legal protection it affords to the right to freedom of speech. By contrast, legal measures to address social media disinformation and misinformation have recently been developed in other legal jurisdictions, such as the EU and the UK. As a result, it seems unlikely that Meta will export its US approach regarding fact-checking to these jurisdictions. Nevertheless, social media users in jurisdictions such as Australia and beyond will have to wait and see as to whether they be subject to Meta's 'Community Notes'. ■

Endnotes

1. Joel Kaplan 'More Speech and Fewer Mistakes' <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>.
2. Ibid.
3. Kate Conger, 'Twitter Has Labelled 38% of Trump's Tweets Since Tuesday' *New York Times* (online, 5 November 2020) <https://www.nytimes.com/2020/11/05/technology/donald-trump-Twitter.html>.
4. Trump was also banned from X (formerly Twitter).
5. Hadas Gold and Liam Reilly, 'Disinformation Experts Blast Trump's Executive Order on Government Censorship as "Direct Assault on Reality"', *CNN Business* (Online, January 23 2025) <https://edition.cnn.com/2025/01/22/media/trump-censorship-executive-order-disinformation/index.html>.
6. Ibid.
7. *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act) [2022] OJ L 277 27/10 1* (Digital Services Act of 2022).
8. Online Safety Act 2023 (UK).
9. See, Communications Legislation Amendment (Combatting Misinformation and Disinformation) Bill (Cth) 2024.
10. Ronald J Krotoszynzki, Jr. 'Disinformation, Misinformation and Democracy: Defining the Problem, Identifying Potentially Effective Solutions, and the Merits of Using a Comparative Legal Approach' in Ronald J Krotoszynzki, Jr., Andras Koltay and Charlotte Garden (eds) *Disinformation, Misinformation and Democracy: Legal Approaches in a Comparative Context* (Cambridge University Press, 2024) 15, citing *United States v Alvarez* 567 US 709 (2012).
11. See eg, Paolo Lobba, 'Holocaust Denial Before the European Court of Human Rights: Evolution of an Exceptional Regime' (2015) 26(1) *European Journal of International Law* 237.
12. See, https://www.oversightboard.com/.
13. Evelyn Douek, 'Meta's Oversight Board: Move Fast with Stable Infrastructure and Humility' (2019) 21(1) *North Carolina Journal of Law & Technology* 48.
14. Stephen Dreyer et al 'Disinformation: Risks, Regulatory Gaps and Countermeasures' (Expert Opinion Commissioned by the Landesanstalt fur Medien, 9 November 2021) 6.
15. European Union, 'High Level Expert Group on Fake News and Online Disinformation' (Final Report, 12 March 2018) <https://digital-strategy.ec.europa.eu/en/library/final-report-high-level-expert-group-fake-news-and-online-disinformation> 10.
16. Margarita Franklin and Mike Torrey, Adversarial Threat Report, Third Quarter 2024, *Meta* (Report, December 4 2024) 3.
17. Ibid.
18. Ibid.
19. Meta, 'Our Approach to Misinformation' <https://transparency.meta.com/en-gb/features/approach-to-misinformation/>.
20. Chris Marsden, Ian Brown and Michael Veale, 'Responding to Disinformation: 10 Recommendations for Regulatory Action and Forbearance' in Martin Moore and Damien Tambini (eds), *Regulating Big Tech: Policy Responses to Digital Dominance* (Oxford University Press, 2022) 201.
21. Ibid.
22. Ibid.
23. Meta, 'More Speech, Fewer Mistakes' <https://about.fb.com/news/2025/01/meta-more-speech-fewer-mistakes/>.
24. As revealed by Frances Haughen's 'Facebook Files': See e.g., Lisa Visentin, 'Stop Trusting Meta' Whistle-blower Francis Haughen Tells Australian MP's', *Sydney Morning Herald* (online, October 21 2021) <https://www.smh.com.au/politics/federal/stop-trusting-Meta-whistleblower-frances-haugen-tells-australian-mps-20211021-p591v9.html>.
25. Meta, Transparency Center, 'Community Notes: A New Way to Add Context to Posts' (10 Feb 2025) <https://transparency.meta.com/en-gb/features/community-notes/>.
26. Ibid.
27. For some interesting research on political bias and X's community notes program, see Sami Nenno, 'Do Community Notes Have a Party Preference?' *Alexander Von Humboldt Institute for Internet and Society, Digital Society Blog* (Blog Post, 20 February 2025) <https://www.hiig.de/en/community-notes-analysis/?utm_source=mailpoet&utm_medium=email&utm_source_platform=mailpoet&utm_campaign=Quarterly%20Newsletter%20Mai%202024>.
28. Ibid.
29. Ibid.
30. Ibid.
31. Mike Isaac and Maggie Haberman, 'Meta Agrees to Pay Trump $40 Million to Settle Lawsuit' *The Sydney Morning Herald* (Online, January 30 2025) <https://www.smh.com.au/business/companies/meta-agrees-to-pay-trump-40-million-to-settle-his-lawsuit-20250130-p5l8an.html>.